

## **Public Use Microdata Sample (PUMS)**

### **Accuracy of the Data (2013-2017)**

## **INTRODUCTION**

This 5-year public use microdata sample (PUMS) for 2013-2017 is a subset of the 2013-2017 American Community Survey (ACS) and Puerto Rico Community Survey (PRCS) samples. It contains the same sample as the combined PUMS 1-year files for 2013, 2014, 2015, 2016, and 2017. Unless otherwise specified, the term “ACS” in this document will refer to both the ACS and PRCS.

This 2013-2017 ACS 5-year PUMS contains five years of data for housing units (HUs) and the population from households and the group quarters (GQ) population. The GQ population, housing units and population from households are all weighted to agree with the ACS counts, which are an average over the five year period (2013-2017). The ACS sample was selected from all counties across the nation, and all municipios in Puerto Rico.

Estimates from the PUMS file are expected to be different from the previously released ACS estimates because they are subject to additional sampling error and further data processing operations. The additional sampling error is a result of selecting the PUMS housing and person records through an additional stage of sampling. In the public use file, the basic unit is an individual housing unit, except for the sample from GQs. For the GQ sample, the basic unit is the person. The population sample is defined as all persons living in households selected in the housing unit sample, plus the persons selected from the GQ sample. Note that microdata records in this sample do not contain names, addresses, or any information that can identify a specific housing unit, GQ or person.

Users of the 2013-2017 ACS 5-year PUMS file can find detailed information on differences between the 2013-2017 files and previous PUMS files in the PUMS ReadMe document. The PUMS ReadMe document for this PUMS file can be found at:

<https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html/>.

## Table of Contents

<b>INTRODUCTION.....</b>	<b>1</b>
<b>CONFIDENTIALITY OF THE DATA .....</b>	<b>3</b>
TITLE 13, UNITED STATES CODE .....	3
DISCLOSURE AVOIDANCE .....	3
DATA SWAPPING.....	3
SYNTHETIC DATA .....	4
PUMAS .....	4
ADDITIONAL MEASURES.....	4
<b>SAMPLE DESIGN.....</b>	<b>4</b>
SAMPLE DESIGN FOR HOUSING UNITS .....	5
SAMPLE DESIGN FOR GROUP QUARTERS .....	6
<b>WEIGHTING .....</b>	<b>6</b>
GROUP QUARTERS PERSON WEIGHTING .....	6
HOUSING UNIT AND HOUSEHOLD PERSON WEIGHTING .....	7
<b>ESTIMATION.....</b>	<b>9</b>
<b>ERRORS IN THE DATA.....</b>	<b>11</b>
SAMPLING ERROR .....	11
NONSAMPLING ERROR .....	11
<b>MEASURING SAMPLING ERROR.....</b>	<b>12</b>
STANDARD ERROR .....	12
CONFIDENCE INTERVALS .....	12
LIMITATIONS.....	13
APPROXIMATING STANDARD ERRORS WITH REPLICATE WEIGHTS .....	14
APPROXIMATING GENERALIZED STANDARD ERRORS WITH DESIGN FACTORS.....	16
EXAMPLES OF STANDARD ERROR CALCULATIONS USING GENERALIZED STANDARD ERROR FORMULAS.....	21
<b>WORKING WITH DOLLAR AMOUNTS.....</b>	<b>24</b>
ADJUSTMENT FACTORS ON THE PUMS FILE.....	24
COMPARING PUMS FILES FROM DIFFERENT PERIODS .....	25
<b>REFERENCES.....</b>	<b>25</b>
<b>PUMAS AFFECTED BY TEL SUPPRESSION.....</b>	<b>26</b>

## CONFIDENTIALITY OF THE DATA

The Census Bureau has implemented a series of steps to protect the confidentiality of the data. Title 13 of the United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual's data can be identified.

The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases.<sup>1</sup> A checklist approach is used to ensure that all potential risks to the confidentiality of the data are considered and addressed.

### Title 13, United States Code

Title 13 of the United States Code authorizes the Census Bureau to conduct censuses and surveys. Section 9 of the same Title requires that any information collected from the public under the authority of Title 13 be maintained as confidential. Section 214 of Title 13 and Sections 3559 and 3571 of Title 18 of the United States Code provide for the imposition of penalties of up to five years in prison and up to \$250,000 in fines for wrongful disclosure of confidential census information.

### Disclosure Avoidance

Disclosure avoidance is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify an individual that has provided information under a pledge of confidentiality. For data tabulations, the Census Bureau uses disclosure avoidance procedures to modify or remove the characteristics that put confidential information at risk for disclosure.

### Data Swapping

Data swapping is a method of disclosure avoidance designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases when creating a table. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics (such as the same number of adults and same number of children). Because the swap often occurs within a neighboring area, there is no effect on the marginal totals for the area or for totals that include data from multiple areas. Because of data swapping, users should not assume that tables with cells having a value of one or two reveal information about specific

---

<sup>1</sup> The Census Bureau's Disclosure Review Board approved the 2013-2017 PUMS 5-year data for release with DRB Clearance number CDDRB-FY19-071.

individuals. Data swapping procedures were first used in the 1990 Census, and were used again in the 2000 Census and the 2010 Census.

## Synthetic Data

The goals of using synthetic data are the same as the goals of data swapping, namely to protect the confidentiality in tables of frequency data. Persons are identified as being at risk for disclosure based on certain characteristics. The synthetic data technique then models the values for another collection of characteristics to protect the confidentiality of that individual.

## PUMAs

The Census Bureau takes further steps to prevent the identification of specific individuals, households, or housing units. The main disclosure avoidance method used is to limit the geographic detail shown in the files. The smallest geographic unit that is identified is the Public Use Microdata Area (PUMA), which is based on a population size of initially around 100,000 or more. No geography smaller than the PUMA can be identified on a PUMS file.

## Additional Measures

Other disclosure avoidance measures used in the PUMS files includes top-coding, age perturbation, weight perturbation and collapsing of detail for categorical variables. The answers to open-ended questions, where an extreme value might identify an individual are top-coded (or bottom-coded). Top-coding (and bottom-coding) substitutes the value of extreme cases with the mean of the highest (or lowest) cases. Top-coded questions include age, income and housing unit value. Age perturbation disguises original data by randomly adjusting the reported ages for a subset of individuals. Weight perturbation disguises the probability of selection for some records. Users should exercise caution when forming estimates near top-coded or bottom-coded values. More information on the variables that receive top or bottom coding in the 2017 PUMS can be found at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

## SAMPLE DESIGN

The 2013-2017 ACS 5-year PUMS sample is the same sample found in each of the 1-year PUMS files for the years 2013, 2014, 2015, 2016, and 2017. It contains five percent of the housing units and five percent of the GQ persons plus some imputed GQ persons in the United States, District of Columbia, and Puerto Rico weighted to represent the average population during five years.

The PUMS GQ sample for data years 2013, 2014, 2015, 2016, and 2017 contained additional imputed records to represent the not-in-sample GQs, which effectively double the total number of records from those years.

See the Accuracy of the Data for the 2017 PUMS 1-year for a further explanation of the PUMS sampling of these imputed GQ records. By including these imputed records in the 2013-2017 ACS 5-year PUMS, the PUMS will agree better with the 2013-2017 5-year full sample ACS for population totals by state and PUMA. More details about the methodology of the large-scale whole person imputation into not-in-sample GQ facilities can be found in the 2017 ACS 1-year Accuracy of the Data at: <https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html>.

## Sample Design for Housing Units

The sampling for HUs (and persons from HUs) was performed independently on the ACS samples of HUs for each of the years 2013, 2014, 2015, 2016 and 2017 as follows:

1. Records of HUs were sorted within each state by: PUMA, ACS weighting area, interview mode, type of vacant, tenure, building type, household type, householder demographics (race, Hispanic origin, sex and age), county, tract, and housing unit weight.
2. Systematic sampling was applied to ACS HUs as described below:
  - a. Within each state, a random number was chosen between zero and the sampling interval. A counter was initialized with the random number.
  - b. At each record, the value of the counter was incremented by one and compared to the sampling interval.
    - i. If the counter's new value was greater than the sampling interval, the HU record was selected for the PUMS and a flag was set to 1. The counter was decreased by the sampling interval with the new value passed to the next record.
    - ii. If the counter was less than the sampling interval, the HU record was not selected for the PUMS and the value of the counter was passed to the next record without altering its value.
3. All HUs selected for PUMS were placed in the PUMS HU sample file.

The PUMS HU sample file was matched to the ACS sample of persons. All persons in selected HUs were placed in the PUMS person sample.

The 2013-2017 5-year ACS Housing Unit estimates for all states may be found on American FactFinder:

[https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17\\_5YR/B25001/0100000US.04000](https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B25001/0100000US.04000).

The PUMS HU sample size may be found in the 2013-2017 ACS 5-year PUMS Record Counts file located on the PUMS Technical Documentation page (<https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>).

## Sample Design for Group Quarters

The sampling for PUMS GQ persons was originally performed on the ACS sample of GQ persons for each of the years 2013, 2014, 2015, 2016 and 2017 as follows:

1. GQ persons were sorted within each state by the size of their GQ facility (large vs small), the type of GQ facility, PUMA, demographics (race, Hispanic origin, sex and age), county, tract, and GQ person weight.
2. Systematic sampling was applied as described above under HUs.
3. All selected GQ persons were added to the PUMS person sample. All imputed records derived from the selected record were also placed in the PUMS person sample.
4. A placeholder record was also placed in the PUMS HU file for each PUMS GQ person record.

The 2013-2017 ACS 5-year estimates for Group Quarters may be found at:

[https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17\\_5YR/B26001/0100000US.04000](https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B26001/0100000US.04000).

The PUMS estimates for Group Quarters may be found in the PUMS Estimates for User Verification on the PUMS Technical Documentation page: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

## WEIGHTING

### Group Quarters Person Weighting

The procedure used to assign the weights to the GQ persons is performed independently within state. The steps are as follows:

#### Initial Weight for GQ Persons

The 5-year PUMS initial weight is the product of the 1-year ACS unrounded weights for the record divided by five and the PUMS subsampling factor. For 2012 and later records, each imputed record received the same subsampling factor as its donor interview. Note that for these data, the ACS weights for sample and imputed records added together represent the GQ universe.

## GQ Person Weighting Factors

### *GQ Person Post-stratification Factor*

This factor adjusts the GQ person weights so that the weighted sample counts equal the published ACS estimates at the state level. The GQ imputed records are not distinguished from the GQ sample records when forming the cells used for the GQ Person Post-stratification Factor Adjustment. Since this adjustment is done at the state level and since noise is added for disclosure avoidance reasons, **only state level PUMS GQ person estimates** will agree with published ACS estimates. This adjustment uses the following groups:

State  $\times$  Institutional/noninstitutional  $\times$  Sex  $\times$  Age Category

### Rounding for GQ Person Weights

The final GQ person weight is rounded to an integer. Rounding is performed so that the sum of the rounded weights is within one person of the sum of the ACS total GQ person estimate for the state.

## Housing Unit and Household Person Weighting

The estimation procedure used to assign the HU and person weights is performed independently within each **PUMA**.

### Initial Weight for Persons and HUs

The 5-year PUMS initial weight is equal to the product of the ACS 1-year final weight for the record and the PUMS subsampling factor divided by five.

### Person Weighting Factors

The person weights are adjusted to agree better with ACS published estimates for householders, spouses, race, Hispanic origin, sex and age by a series of two steps that are repeated until a stopping criterion is met. This is an iterative proportional fitting or raking process. The person weights are individually adjusted at each step as described below.

The two steps are as follows:

### *Spouse Equalization/Householder Equalization Raking Factor*

This factor is applied to individuals based on the combination of their status of being in a married-couple or unmarried-partner household and whether they are the householder. All persons are assigned to one of four groups:

---

1. Householder in a married-couple or unmarried-partner household
2. Spouse or unmarried partner in a married-couple or unmarried-partner household (non-householder)
3. Other householder
4. Other non-householder

The weights of persons in the first two groups are adjusted so that their sums are each equal to the ACS estimate of married-couple or unmarried-partner households using the ACS housing unit weight. The weights of persons in the third group are adjusted so that the sum is equal to the ACS estimate of occupied housing units not having a partner using the housing unit weight. The weights of persons in the fourth group are adjusted to agree with the ACS total population minus the first three groups. The goal of this step is to produce more consistent estimates of spouses or unmarried partners and married-couple and unmarried-partner households while simultaneously producing more consistent estimates of householders, occupied housing units, and households.

#### *Demographic Raking Factor*

This factor is applied to individuals based on their age, race, sex and Hispanic origin. It adjusts the person weights so that the weighted sample counts equal ACS population estimates by age, race, sex, and Hispanic origin at the PUMA level. Because of collapsing of groups in applying this factor, only total population is assured of agreeing precisely with the published ACS 2013-2017 population estimates at the PUMA level.

This uses the following groups within each PUMA (note that there are 13 Age groupings):

Race / Ethnicity (non-Hispanic White, non-Hispanic Black, non-Hispanic American Indian or Alaskan Native, non-Hispanic Asian, non-Hispanic Native Hawaiian or Pacific Islander, and Hispanic (any race))  $\times$  Sex  $\times$  Age Groups

These two steps are repeated several times until the estimates at the PUMA level achieve their optimal consistency with regard to the spouse and householder equalization. The final Person Weighting Factor is then equal to the product of the factors from all of the iterations of these two adjustments. The unrounded person weight is then equal to the product of Person Weighting Factor times the initial person weight.

#### **Rounding of Person Weights**

The person weight after the Person Weighting Factor has been applied is rounded to an integer. Rounding is performed so that the sum of the rounded weights is within one person of the sum of the ACS total person from HU's estimates within state and PUMA.

## **Householder Adjustment Factor (HHRF)**

This factor applied to occupied housing units is the same as the Person Weighting factor from the person weighting. After this stage, the weight of the housing unit is identical to the unrounded person weight of the householder after the Person Weighting Factor is applied.

## **Housing Unit Control factor**

This factor adjusts PUMS housing unit estimates to agree with the published ACS housing unit estimates for housing units with married couples (or partners), occupied housing units without partners and vacant housing units.

## **Rounding of Housing Unit weights**

The Housing Unit weight after the Housing Unit Control Factor is applied is rounded to an integer. Rounding is performed so that the sum of the rounded weights is within one housing unit of the sum of the ACS total HU's estimates within state and PUMA.

For a detailed description of how the original ACS weights are computed, see the 2013-2017 ACS Multiyear Accuracy of the Data at: <https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html>

## **ESTIMATION**

To produce estimates or tabulations of characteristics from the PUMS, add the weights of all persons or HUs that possess the characteristic of interest.<sup>2</sup> For instance, if the characteristic of interest is “total number of black teachers”, determine the race and occupation of all persons and cumulate the weights of those who match the characteristics of interest. To get estimates of proportions, divide the weighted estimate of persons or HUs with a given characteristic by the weighted estimate of the base. For example, the proportion of “black teachers” is obtained by dividing the weighted estimate of black teachers by the estimate of teachers.

PUMS estimates are expected to be different from published ACS estimates that are based on the full set of data because of the additional sampling. The exception will be characteristics controlled by the ratio-estimate factors.

---

<sup>2</sup> Users should exercise caution when forming estimates near top-coded or bottom-coded values. More information on the variables that receive top or bottom coding in the PUMS can be found at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

Note that the housing unit file contains some records with blank weights. These are the GQ placeholder records.<sup>3</sup> The housing unit weights were set to zero for these records since they are not housing units, but persons. For confidentiality reasons, the GQ data are not provided at the level of an address but only at the person-level. All of the GQ person data are included in the PUMS person file except variable for the “Yearly food stamp/Supplemental Nutrition Assistance Program (SNAP) recipiency” (FS), which are the only data, included on the GQ placeholder records in the housing unit file. For food stamp recipiency estimates of persons in GQs, you will need to match the placeholder records to the person file to get the person weights.

A note to GQ data users. There are limitations to the usefulness of GQ estimates at the PUMA level. The PUMS weighting controls the GQ estimates to agree with the ACS state level estimates. Depending on the application or analysis, GQ data users should consider working with state level estimates rather than PUMAs.

In a limited number of geographies, the ACS PUMS file has suppressed variables affecting PUMS data. The suppression was due to nonsampling error or issues with interpreting the recode. Three variables were affected. In 2015, the telephone variable (TEL) was suppressed in five PUMAs. In 2016, the telephone variable (TEL) was suppressed in fourteen PUMAs. In these cases, the nonsampling errors could not be edited or corrected before the publication of the PUMS file. Specific PUMAs with suppressed data for TEL are listed in at the end of this document. In addition, beginning in 2012, the complete plumbing facilities recode (PLM) was suppressed in Puerto Rico.

Within the PUMAs listed in the user note, the variable TEL were assigned the value ‘8’ to identify it as suppressed due to data problems. In order to estimate the telephone rate for the states impacted by the suppressed values, use the estimate having the value of ‘1’ divided by the sum of the estimate having the value of ‘1’ and the estimate having the value of ‘2’. In the examples shown, the fertility rate is the number out of 1000 and the telephone rate is a percent:

$$\text{Telephone Rate} = 100 * \frac{\sum \text{pwgtp for all cases having TEL} = '1'}{\sum \text{wgtp for all cases having TEL} = '1' \text{ or } '2'}$$

The variable PLM was assigned the value of ‘9’ in all PUMAs in Puerto Rico to mean “not applicable”. See the table at the end of this document for a list of PUMA codes affected by the TEL variables.

---

<sup>3</sup> To identify HU and GQ placeholder records on the PUMS housing file, see the TYPE variable in the PUMS data dictionary: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

In 2016, the questions pertaining to business on property (BUS) and presence of a flush toilet (TOIL) were removed from the ACS questionnaire. As such, there is not data for data year 2016 for variables BUS and TOIL on the 2013-2017 ACS 5-year PUMS file. Rather, they are each assigned a value of '9' for data year 2016 cases. Likewise, the corresponding allocation flags, FBUSP and FTOILP, were assigned a value of zero on the 2013-2017 file.

Data users should use caution when dealing with the variables BUS, TOIL, FBUSP and FTOILP. Five year estimates for these variables cannot be derived from the 2013-2017 ACS 5-year PUMS file, as only four years' worth of data are provided. Any five year estimates derived for BUS, TOIL, FBUSP and FTOILP from the 2013-2017 file will be inaccurate.

These variables were retained on the file as they were previously included as components of the variables SVAL and PLM, respectively. The removal of the variables caused a change in the components of these variables between data years 2015 and 2016. For 2015, BUS was included as a component in SVAL; for 2016 and later, SVAL does not consider the presence of a business on the property. Likewise, PLM required a flush toilet in 2015; in 2016 and later, complete plumbing is not defined by TOIL. Keeping the 2013, 2014, and 2015 data values on the 2013-2017 ACS 5-year PUMS files allows users to create recodes of PLM and SVAL that will be comparable across all five data years. For more information, please reference the BUS and TOIL User Note at: <https://www.census.gov/programs-surveys/acs/technical-documentation/user-notes.html>.

## **ERRORS IN THE DATA**

Every sample survey is subject to two types of error: sampling error and nonsampling error.

### **Sampling Error**

The data in the ACS products are estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology. The estimates from the chosen sample also differ from other samples of HUs and persons within those HUs. Sampling error in data arises due to the use of probability sampling, which is necessary to ensure the integrity and representativeness of sample survey results. The implementation of statistical sampling procedures provides the basis for the statistical analysis of sample data.

Estimates made with PUMS data are subject to additional sampling error because the PUMS data consists of a subset of the full ACS sample. Thus standard errors of PUMS estimates can be larger than standard errors that would be obtained using all of the ACS data.

### **Nonsampling Error**

In addition to sampling error, data users should realize that other types of errors may be introduced during any of the various complex operations used to collect and process survey

data. For example, operations such as data entry from questionnaires and editing may introduce error into the estimates. These and other sources of error contribute to the nonsampling error component of the total error of survey estimates. Nonsampling errors may affect the data in two ways. Errors that are introduced randomly increase the variability of the data. Systematic errors, which are consistent in one direction, introduce bias into the results of a sample survey. The Census Bureau protects against the effect of systematic errors on survey estimates by conducting extensive research and evaluation programs on sampling techniques, questionnaire design, and data collection and processing procedures. In addition, an important goal of the ACS is to minimize the amount of nonsampling error introduced through nonresponse for sample HUs. One way of accomplishing this is by following up on mail nonrespondents during the CATI and CAPI phases.

More information about the control of nonsampling error can be found in ACS Multiyear Accuracy of the Data (2013-2017) at: <https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html>.

## MEASURING SAMPLING ERROR

### Standard Error

Standard Error is a measure of the deviation of a sample estimate from the average of all possible samples. Sampling error and some types of nonsampling error are estimated by the standard error. The sample estimate and its estimated standard error permit the construction of interval estimates with a prescribed confidence that the interval includes the average result of all possible samples.

Two methods are provided for calculating the standard errors of PUMS estimates: a Successive Difference Replicate (SDR) method (using replicate weights) and a Generalized Variance Function method (using design factors). Replicate weights have been provided with the ACS PUMS files since the 2005 PUMS. Design factors (a type of generalized variance function) is a method used by the Census 2000 PUMS and also in use by the ACS PUMS since 2000. **It is important to keep in mind that there will be differences between the standard error approximations computed by these two methods.** Using the replicate weights will produce a more accurate estimate of a standard error.

### Confidence Intervals

A sample estimate and its estimated standard error may be used to construct confidence intervals about the estimate. These intervals are ranges that will contain the average value of the estimated characteristic that results over all possible samples, with a known probability.

For example, if all possible samples that could result under the PUMS sample design were independently selected and surveyed under the same conditions, and if the estimate and its estimated standard error were calculated for each of these samples, then:

Approximately 68 percent of the intervals from one estimated standard error below the estimate to one estimated standard error above the estimate would contain the average result from all possible samples;

Approximately 90 percent of the intervals from 1.645 times the estimated standard error below the estimate to 1.645 times the estimated standard error above the estimate would contain the average result from all possible samples.

Approximately 95 percent of the intervals from two estimated standard errors below the estimate to two estimated standard errors above the estimate would contain the average result from all possible samples.

These intervals are referred to as 68 percent, 90 percent, and 95 percent confidence intervals, respectively. An example of how to construct a 90 percent confidence interval follows:

Add and subtract 1.645 times the standard error to the estimate to yield the lower and upper bounds of a 90% confidence interval around the estimate (EST).

$$LB = \text{Lower bound} = EST - 1.645 * SE(EST)$$

$$UB = \text{Upper bound} = EST + 1.645 * SE(EST)$$

The 90% confidence interval is the interval (LB, UB).

## **Limitations**

The user should be careful when computing and interpreting standard errors and confidence intervals.

### **Nonsampling Error**

The estimated standard errors included in this data product do not include all portions of the variability due to nonsampling error that may be present in the data. In particular, the standard errors do not reflect the effect of correlated errors introduced by interviewers, coders, or other field or processing personnel. Nor do they reflect the error from imputed values due to missing responses. Thus, the standard errors calculated represent a lower bound of the total error. As a result, confidence intervals formed using these estimated standard errors may not meet the stated levels of confidence (i.e., 68, 90, or 95 percent). Thus, some care must be exercised in the interpretation of the data in this data product based on the estimated standard errors.

## Very Small (Zero) or Very Large Estimates

The value of almost all PUMS characteristics is greater than or equal to zero by definition. For zero or small estimates, use of the method given previously for calculating confidence intervals relies on large sample theory, and may result in negative values that, for most characteristics, are not admissible. In this case the lower limit of the confidence interval is set to zero by default. A similar caution holds for estimates of totals close to a control total or estimated proportions near one, where the upper limit of the confidence interval is set to its largest admissible value. In these situations the level of confidence of the adjusted range of values is less than the prescribed confidence level.

## Approximating Standard Errors with Replicate Weights

Replicate weights can be used to calculate what are referred to as successive difference replicate (SDR) or direct standard errors. Standard errors for the published ACS tabulations are calculated using the SDR method. Direct standard errors will often be more accurate than generalized standard errors, although they may be more inconvenient for some users to calculate. The advantage of using SDR method is that a single formula is used to calculate the standard error of many types of estimates.

Each PUMS housing unit and person record contains 80 replicate weights. These replicate weights were formed from the ACS replicate weights adjusted for PUMS subsampling and ratio adjustments. For any estimate  $X$ , 80 replicate estimates are also computed using the replicate weights. For this discussion, we refer to  $X$  as the ‘full sample estimate.’ The first replicate estimate,  $X_1$ , is computed using the first replicate weight, the second replicate estimate,  $X_2$ , is computed using the second replicate weight, and so on. Each replicate estimate is computed using the replicate weights in the same way that the full sample estimate  $X$  is computed.

**NOTE:** When programming the replicate weight standard errors, users will find the eighty replicate weights can be positive, zero or negative. The negative replicate weights are partly due to the addition of the Group Quarters (GQ) population to the full ACS weighting process. Within a weighting cell, GQ estimates were subtracted from population totals, sometimes resulting in negative values for the cell. The cells were collapsed in such a way as to prevent a final cell from being zero or negative for the full sample weights. The full sample weights are never negative. This restriction was not placed on the replicate weights since their only purpose is to represent the variability of the sample. PUMS replicate weights are based on ACS replicate weights so negative values may occur. Keep in mind that the replicate weights are only to be used to estimate the variance with the formula provided in the PUMS accuracy document.

The standard error of X can be computed after the replicate estimates  $X_1$  through  $X_{80}$  are computed. The standard error is estimated using the sum of squared differences between each replicate estimate  $X_r$  and the full sample estimate X. The standard error formula is:

$$SE(X) = \sqrt{\frac{4}{80} \sum_{r=1}^{80} (X_r - X)^2}$$

If X is zero, then use the generalized variance method for zero estimates given the Standard Errors for Totals and Percentages section of this document, to approximate the standard error.

Data users who wish to see worked examples may consult the documentation for the ACS Variance Replicate Tables, located here: <https://www.census.gov/programs-surveys/acs/technical-documentation/variance-tables.html>.

The standard error can be used to form a 90% confidence interval around the estimate (X) as follows:

$$LB = \text{Lower bound} = X - 1.645 * SE(X)$$

$$UB = \text{Upper bound} = X + 1.645 * SE(X)$$

The 90% confidence interval is the interval (LB, UB).

As previously mentioned, we consider the replicate weight SEs to be more accurate than the design factor SEs. For exceptions, please note the following:

After using replicate weight SEs, some users may notice that occasionally the SE is zero for an estimate. The user may want to know if this is accurate. Except for controlled estimates, all PUMS estimates are based on a sample of the population and should not have a SE of zero. However, if the estimate is a controlled count (or total) such as total population or total GQ population in a state, there is no sampling variability in the estimate. It is expected that the replicate weight SE and MOE will be zero for some controlled estimates.

If your estimate is a median, the replicate weight method may yield a SE of zero. This occurs when several records in the middle of the distribution were rounded to the same value, or when the characteristic contains few records, such as a median based on less than five records. Rounding by respondents, as well as rounding by PUMS edits may mask the variability in the median. In order to yield a more adequate standard error for that case, use the design factor method to estimate the SE of a median.

Examples of PUMS estimates with replicate weight standard errors are found in the document PUMS Estimates for User Verification at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

## Approximating Generalized Standard Errors with Design Factors

### Note on the Design Factors

Note that beginning in 2017, the design factors are no longer included in this document. They are published in a comma separated value (CSV) file located at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

### Totals and Percentages

The design factors provided in comma-separated value (CSV) file entitled “2013-2017 ACS 5-year PUMS Design Factors (Attachment A)” can be used to approximate the standard errors of most sample estimates of *totals* and *proportions*. Design factors are given by subject for the United States, all 50 states, the District of Columbia, and Puerto Rico. The term "subject" refers to a characteristic, such as age for persons and tenure for HUs. The design factors reflect the effects of the actual sample design and estimation procedures used for the ACS. To approximate the standard error for most estimates, use the following formulas:

#### Total Formula

$$SE(Y) = DF \times \sqrt{\frac{95}{5} \times Y \left(1 - \frac{Y}{N}\right)}$$

Where:

DF = Design Factor

N = Size of Population in the Geographic Area

Y = Estimate of Characteristic Total

#### Percent Formula

$$SE(p) = DF \times \sqrt{\left(\frac{95}{5B}\right) \times p(100 - p)}$$

Where:

DF = Design Factor

B = Denominator of Estimated Percentage

p = Estimated Percentage

The values of N and the design factor can be determined as follows:

1. For the value of N, obtain the number of persons, number of households or number of HUs, respectively for the geographies you are interested in. If the estimate is of HUs then use the number of HUs; if the estimate is of families or households then use the number of households; otherwise use the number of persons.
2. Select the appropriate table from the comma separated value (CSV) PUMS Design Factors (Attachment A) file, located at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>. Use the design factor for the United States when estimating characteristics for the United States or geographic areas that cover more than one state. Use the table for a specific state when estimating characteristics for that state or geographic areas that are contained entirely within that state.
3. Then use the selected characteristic to obtain the appropriate design factor for the characteristic; for example, educational attainment or ancestry. If the estimate is a combination of two or more characteristics, we suggest the following guideline: Use the largest design factor for this combination of characteristics. **The only exception to this is for items crossed with race or Hispanic Origin. For an item(s) crossed with race or Hispanic Origin, use the largest design factor not including the race or Hispanic Origin design factor.**

An inspection of the formulas used to calculate the simple random sampling standard errors suggests that when dealing either with zero estimates or with very small estimates of totals and percentages, the standard error estimates approach zero. This is also the case for very large estimates of totals and percentages. Zero or small estimates, like any other sample estimates, are still subject to sampling variability and therefore an estimated standard error of zero or close to zero is not adequate. Use the recommended procedures below for estimates that fit the following descriptions:

1. **An estimated total is less than 425 or within 425 of the total size of the tabulation area.** Use a basic standard error of 110 multiplied by the design factor for the type of estimate.
2. **An estimated percentage is less than 2 or greater than 98.** Use a value of 2 for the estimated percentage in the percent formula.
3. **The denominator of a percentage is zero.** There are no sample observations available to compute an estimate of a proportion or an estimate of its standard error.

## Sums and Differences

For the sum or difference between two estimates, the standard error is approximately the square root of the sum of the two individual standard errors squared:

$$SE(\hat{X} + \hat{Y}) = SE(\hat{X} - \hat{Y}) = \sqrt{[SE(\hat{X})]^2 + [SE(\hat{Y})]^2}$$

This method is, however, an approximation as the two estimates of interest in a sum or a difference are likely to be correlated. If the two quantities X and Y are positively correlated, this method underestimates the standard error of the sum of X and Y and overestimates the standard error of the difference between the two estimates. If the two estimates are negatively correlated, this method overestimates the standard error of the sum and underestimates the standard error of the difference.

## Ratios

Frequently, the statistic of interest is the ratio of two variables, where the numerator is not a subset of the denominator. An example is the ratio of students to teachers in public elementary schools. The standard error of the ratio between two sample estimates is estimated as follows:

$$SE\left(\frac{\hat{X}}{\hat{Y}}\right) = \left(\frac{\hat{X}}{\hat{Y}}\right) \times \sqrt{\frac{[SE(\hat{X})]^2}{\hat{X}} + \frac{[SE(\hat{Y})]^2}{\hat{Y}}}$$

If the ratio is a proportion, that is, the numerator is a subset of the denominator, then follow the procedure outlined in the Standard Errors for Totals and Percentages section of this document.

## Medians

The sampling variability of an estimated median of a variable depends on the form of the distribution and the size of its base. The standard error of an estimated median is approximated by constructing a 68 percent confidence interval. Estimate the 68 percent confidence limits of a median based on sample data using the following procedure.<sup>4</sup>

1. Obtain the weighted frequency distribution for the selected variable using user defined categorical values. Cumulate these frequencies to yield the base. In general, variables pertaining to income could use a category as small as \$2,500, for example

<sup>4</sup> The design factor method shown here for medians is preferred over the replicate weight method whenever the replicate weight method gives a standard error of zero. This may happen due to having several records in the middle of the range that have exactly the same value. Be aware that PUMS dollar values are rounded to the nearest 100 for values between 1,000 and 50,000 and rounded to the nearest 1,000 above 50,000. This increases the number of respondents with exactly the same value. The amount of rounding done by respondents is unknown, but could be substantial. Since rounding may cause the number of records with exactly the same value to increase, and might cause all 80 replicates to yield the same median, the replicate weight formula can give a standard error of zero. To avoid this, it is possible to calculate the medians using a categorical method with linear interpolation for all 80 replicates, OR simply use the design factor method to estimate the standard errors.

\$0-\$2,499, \$2,500-\$4,999, etc. In Example 3, only sixteen rows are used (for simplicity), which causes the income category widths to be larger than ideal. Other variables such as gross rent should use smaller category widths than the income variables.

2. Determine the standard error of a 50 percent proportion using the formula in the Standard Errors for Totals and Percentages section of this document.

$$SE(50\% \text{ percent}) = DF \times \sqrt{\left(\frac{95}{5B}\right) \times 50^2}$$

3. Subtract from and add to 50 percent the standard error determined in step 2.

$$p_{\text{lower}} = 50 - SE(50\% \text{ percent})$$

$$p_{\text{upper}} = 50 + SE(50\% \text{ percent})$$

4. Determine the categories in the distribution that contain  $p_{\text{lower}}$  and  $p_{\text{upper}}$ .

If  $p_{\text{lower}}$  and  $p_{\text{upper}}$  fall in the same category, follow step 5. If  $p_{\text{lower}}$  and  $p_{\text{upper}}$  fall in different categories, go to step 6.

5. If  $p_{\text{lower}}$  and  $p_{\text{upper}}$  fall in the same category, do the following:

- Define A1 as the smallest value in that category.
- Define A2 as the smallest value in the next (higher) category.
- Define C1 as the cumulative percent of units strictly less than A1.
- Define C2 as the cumulative percent of units strictly less than A2.

Use the following formulas to determine the lower and upper bounds for a confidence interval about the median:

$$\text{Lower Bound} = \left[ \frac{p_{\text{lower}} - C1}{C2 - C1} \right] \times (A2 - A1) + A1$$

$$\text{Upper Bound} = \left[ \frac{p_{\text{upper}} - C1}{C2 - C1} \right] \times (A2 - A1) + A1$$

6. If  $p_{\text{lower}}$  and  $p_{\text{upper}}$  fall in different categories, do the following:

For the category containing  $p_{\text{lower}}$ : Define A1, A2, C1, and C2 as described in step 5. Use these values and the formula in step 5 to obtain the lower bound.

For the category containing p\_upper: Define new values for A1, A2, C1, and C2 as described in step 5. Use these values and the formula in step 5 to obtain the upper bound.

7. Use the lower and upper bounds determined in steps 5 or 6 to calculate the standard error of the median.

$$SE(\text{median}) = 1/2 \times (\text{Upper Bound} - \text{Lower Bound})$$

## Means

A mean is defined here as the average quantity of some characteristic (other than the number of people, HUs, households, or families) per person, housing unit, household, or family. For example, a mean could be the average annual income of females age 25 to 34. The standard error of a mean can be approximated by the formula below. Because of the approximation used in developing this formula, the estimated standard error of the mean obtained from this formula will generally underestimate the true standard error.

$$SE(Y) = DF \times \sqrt{\left(\frac{95}{5B}\right) \times s^2}$$

Where:

B is the base (denominator) of the mean

$s^2$  is the sample variance of the characteristic based on weighted data.

The value of  $s^2$  can be computed using the formula:

$$s^2 = \frac{\sum_{i=1}^n w_i y_i^2 - [\sum_{i=1}^n w_i y_i]^2 / \sum_{i=1}^n w_i}{(\sum_{i=1}^n w_i) - 1}$$

Where:

$w_i$  is the weight of the  $i^{th}$  sample record

$y_i$  is the value of the characteristic for the  $i^{th}$  sample record

$n$  is the number of sample records

Note that  $\sum_{i=1}^n w_i$  is the weighted estimate of persons/HUs in the sample (ex. the number of females age 25 to 34, and  $\sum_{i=1}^n w_i y_i$  is the weighted aggregate estimate for the characteristic of interest (ex. the aggregate income of females age 25 to 34).

## Examples of Standard Error Calculations using Generalized Standard Error Formulas

We will present some examples based on the 2009-2013 PUMS 5-year data to demonstrate the use of the generalized standard error formulas.

### Example 1 – Using Design Factors to Estimate the Standard Error of a Total

The estimated number of people 15 years or over who were never married is 2,136,436 from the PUMS data for the state of Virginia. To calculate the standard error, we use the total formula given in the section Standard Errors for Totals and Percentages. In this formula, Y is our estimate of 2,136,436 and N is the total PUMS population for the state of Virginia, which is 8,256,630. The design factor for “Marital Status” is 1.4.

$$SE = 1.4 \times \sqrt{\frac{95}{5} \times 2,136,436 \times \left(1 - \frac{2,136,436}{8,256,630}\right)} = 7,679.46$$

To calculate the margin of error, simply multiply 7,679.46 by 1.645 to get 12,632.72. To obtain the lower and upper bounds of the 90 percent confidence interval around 2,136,436 using the margin of error, simply add and subtract 12,632.72 from 2,136,436. Thus the 90 percent confidence interval for this estimate is [2,136,436 - 12,632.72] to [2,136,436 + 12,632.72] or 2,123,803.28 to 2,149,068.72.

### Example 2 – Using Design Factors to Estimate the Standard Error of a Proportion or Percentage

The estimated percent of people 25 years or over with a bachelor’s degree or higher in Louisiana is  $22.4190 = 100 * (681,488 / 3,039,780)$  from the PUMS data. To calculate the standard error, we use the percent formula given in the section Standard Errors for Totals and Percentages. Use the denominator of the percentage, 3,039,780 in the formula. The design factor for “Educational Attainment” is 1.5.

$$SE = 1.5 \times \sqrt{\frac{95}{5 \times 3,039,780} \times 22.4190 \times (100 - 22.4190)} = 0.1564$$

To calculate the margin of error, multiply 0.1564 by 1.645 to get 0.2573. To obtain the lower and upper bounds of the 90 percent confidence interval around 22.4190 percent using the margin of error, simply add and subtract 0.2573 from 22.4190. Thus the 90 percent confidence interval for this estimated percentage is [22.4190 - 0.2573] to [22.4190 + 0.2573] or 22.16 to 22.68.

### Example 3 – Calculating the Standard Error of a Median

Users need to form a weighted frequency distribution for the variable of interest. Table 3 below shows one possible weighted frequency distribution for adjusted household income in Massachusetts.

**Table 1: A Possible Distribution Frequency for Adjusted Household Income in MA**

Adjusted Household Income	Frequency	Cumulative Frequency	Cumulative Percent
Less than \$10,000	153,739	153,739	6.03
\$10,000 to \$14,999	130,852	284,591	11.16
\$15,000 to \$19,999	113,550	398,141	15.62
\$20,000 to \$24,999	105,230	503,371	19.74
\$25,000 to \$29,999	95,824	599,195	23.50
\$30,000 to \$34,999	102,957	702,152	27.54
\$35,000 to \$39,999	90,972	793,124	31.11
\$40,000 to \$44,999	87,818	880,942	34.55
\$45,000 to \$49,999	83,793	964,735	37.84
\$50,000 to \$59,999	174,388	1,139,123	44.68
\$60,000 to \$74,999	231,284	1,370,407	53.75
\$75,000 to \$99,999	318,700	1,689,107	66.25
\$100,000 to \$124,999	244,795	1,933,902	75.85
\$125,000 to \$149,999	178,104	2,112,006	82.83
\$150,000 to \$199,999	202,797	2,314,803	90.79
\$200,000 or more	234,913	2,549,716	100.00

The base is the cumulative sum of the weighted frequencies, which is 2,549,716.

Determine the standard error of a 50 percent proportion, using as the denominator the cumulative sum of the weighted frequencies, 2,549,716. For this example, the design factor for household income is 1.5.

$$SE(50 \text{ percent}) = 1.5 \times \sqrt{\frac{95}{5 \times 2,549,716} \times 50^2} = 0.20$$

Calculate  $p_{\text{lower}}$  and  $p_{\text{upper}}$ .

$$p_{\text{lower}} = 50 - SE(50 \text{ percent}) = 49.80$$

$$p_{\text{upper}} = 50 + SE(50 \text{ percent}) = 50.20$$

Determine the categories that contain  $p_{\text{lower}}$  and  $p_{\text{upper}}$ . The first category with a cumulative percentage that is greater than 49.80 is \$60,000 to \$74,999. The first category with a cumulative

percentage that is greater than 50.20 is \$60,000 to \$74,999. Since p\_lower and p\_upper fall in the same category, follow the instructions given in step 5 of the section Standard Errors for Medians.

Define A1, A2, C1, and C2: A1 = 60,000, A2 = 75,000, C1 = 44.68 and C2 = 53.75. Calculate the lower bound and upper bound using these values.

$$\text{Lower Bound} = \left[ \frac{49.80 - 44.68}{53.75 - 44.68} \right] \times (75,000 - 60,000) + 60,000 = 68,467.48$$

$$\text{Upper Bound} = \left[ \frac{50.20 - 44.68}{53.75 - 44.68} \right] \times (75,000 - 60,000) + 60,000 = 69,129.00$$

Finally, calculate the standard error of the median:

$$\text{SE}(\text{median}) = 1/2 \times (69,129.00 - 68,467.48) = 330.76$$

#### Example 4 – Calculating the Standard Error of a Mean

Suppose we wish to estimate mean adjusted person income of females age 25 to 34 in Alabama. Table 4 below summarizes the computation of the terms in the formula for  $s^2$ . The PUMS data for Alabama has 12,893 records for females age 25 to 34 that have a non-missing value for person income.

**Table 2: Computations for Mean Adjusted Person Income of Females Age 25 to 34 in Alabama**

Sample Record	$y_i$	$w_i$	$w_i y_i$	$w_i y_i^2$
1	21,462	5	107,309	2,303,061,466
2	21,462	13	279,004	5,987,959,811
3	9,658	32	309,051	2,984,767,660
4	65,459	21	1,374,633	89,981,762,995
.				
.				
12,893	55,169.65	17	937,884	51,742,728,026
<b>Total</b>	<b>274,162,709</b>	<b>317,090</b>	<b>6,575,359,529</b>	<b>302,151,315,109,878</b>

Note: The inflation-adjusted income shown in column 2 was the person income multiplied by the ADJINC variable. Unrounded adjusted income was used to compute the products in the third and fourth columns.

The mean adjusted income is:

$$\bar{Y} = \frac{6,575,359,529}{317,090} = 20,736.57$$

and  $s^2$  is computed as follows:

$$s^2 = \frac{302,151,315,109,878 - 6,575,359,529^2/317,090}{317,090 - 1} = 522,884,429$$

The design factor for person income in Alabama, is 1.6. The standard error of the mean can now be calculated:

$$SE(\bar{Y}) = 1.6 \times \sqrt{\frac{95}{5 \times 317,090} \times 522,884,429} = 283$$

## WORKING WITH DOLLAR AMOUNTS

Dollar variables must be adjusted into a common year before using them to form estimates. Generally, the older years are adjusted to the most recent year covered by the analysis. Data in the current ACS 5-year PUMS were collected over five years, so the adjustment will inflate 2013, 2014, 2015, and 2016 dollars into 2017 values.

### Adjustment Factors on the PUMS File

The PUMS data dictionary for 2013-2017 describes two adjustment factors on the 5-year file that put dollar values into 2017 dollars:

**ADJINC** – inflation adjustment factor for income variables, such as household income, self-employment income, retirement income and wages.

**ADJHSG** – inflation adjustment factor for most housing dollar variables, such as utility costs, rent, food stamps, and condominium fees.

For more details, see the PUMS data dictionary at <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

For example, multiply the household income variable by ADJINC to adjust household income into 2017 dollars. One reason this adjustment is needed is because interviews in the ACS were conducted throughout the year for a reference period that included twelve previous months. Application of the adjustment factor will convert amounts to 2017 dollars.

For example, multiply ADJHSG times the monthly rent to adjust rent into 2017 dollars. All records get this adjustment, although the records interviewed in 2017 have a factor of 1.

Note that the values of ADJINC and ADJHSG are the same for all sample cases from the same year. This is for disclosure avoidance reasons, that is, so that the month of interview cannot be identified by the adjustment factor.

## Comparing PUMS files from Different Periods

When comparing dollar estimates from the 2013-2017 ACS 5-year PUMS file to estimates from other years, an additional adjustment is necessary to convert the amounts into dollars from a common year (after applying the adjustment factor described in the previous paragraphs). We use the CPI-U-RS adjustment factors from the Bureau of Labor Statistics. These factors can be found in the first table in the PDF file at:

<https://www.bls.gov/cpi/research-series/> [For example, to express year 2000 dollars in terms of 2017 dollars, multiply the 2000 dollars by  $361.0/252.9 = 1.427$ ].

## REFERENCES

[1]. ACS Multiyear Accuracy of the Data (2013-2017)

<https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html>

[2]. Design and Methodology of the American Community Survey:

<https://census.gov/programs-surveys/acs/methodology/design-and-methodology.html>

[3]. PUMS Accuracy of the Data for 1-Year PUMS files:

<https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html/>

[4]. PUMS Data Dictionary:

<https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html/>

[5]. Updated CPI-U-RS, All Items, 1977-2017

<https://www.bls.gov/cpi/research-series/>

**PUMAs Affected by TEL Suppression**

<b>Year</b>	<b>State FIPS Code</b>	<b>PUMA</b>
2015	05	01100
2015	12	10700
2015	12	10900
2015	21	02600
2015	55	01601
2016	17	02100
2016	37	04600
2016	37	04700
2016	37	05100
2016	45	00603
2016	45	00604
2016	48	01901
2016	48	01902
2016	48	01903
2016	48	06801
2016	48	06802
2016	48	06803
2016	48	06804
2016	48	06807